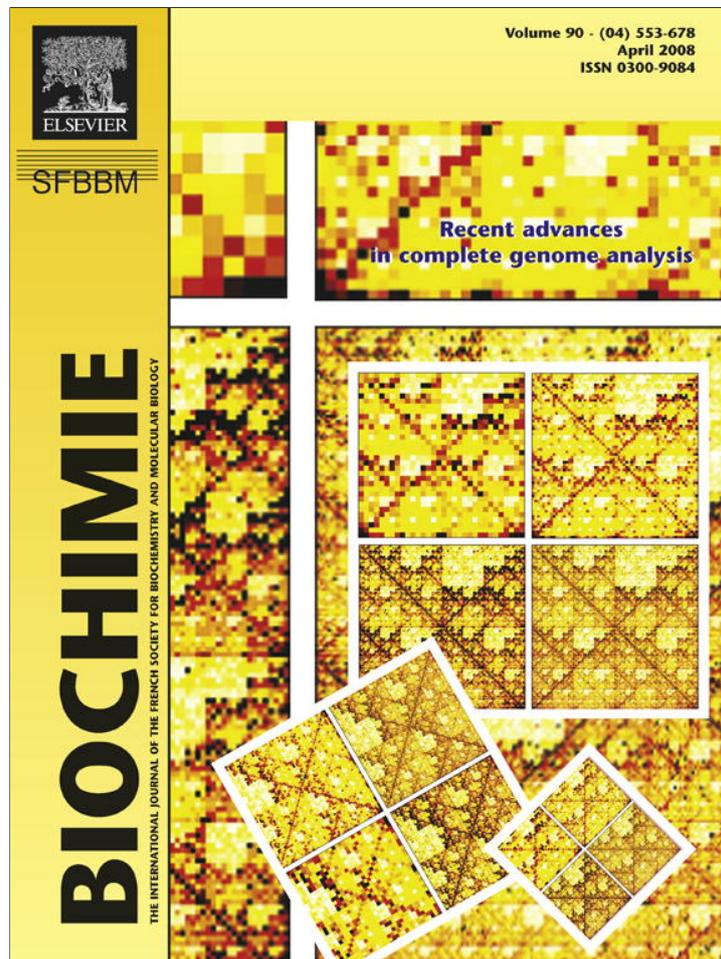


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Research paper

IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm

Patrice Duroux^a, Quentin Kaas^a, Xavier Brochet^a, Jérôme Lane^a,
Chantal Ginestoux^a, Marie-Paule Lefranc^{a,b,*}, Véronique Giudicelli^a

^a *IMGT[®], the international ImMunoGeneTics information system[®], Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France*

^b *Institut Universitaire de France, 103 Bd Saint-Michel, 75005 Paris, France*

Received 8 June 2007; accepted 4 September 2007

Available online 11 September 2007

Abstract

IMGT[®], the international ImMunoGeneTics information system[®] (<http://imgt.cines.fr>), is the reference in immunogenetics and immunoinformatics. IMGT standardizes and manages the complex immunogenetic data which include the immunoglobulins (IG) or antibodies, the T cell receptors (TR), the major histocompatibility complex (MHC) and the related proteins of the immune system (RPI) which belong to the immunoglobulin superfamily (IgSF) and the MHC superfamily (MhcSF). The accuracy and consistency of IMGT data and the coherence between the different IMGT components (databases, tools and Web resources) are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, “IDENTIFICATION”, “DESCRIPTION”, “CLASSIFICATION”, “NUMEROTATION”, “LOCALIZATION”, “ORIENTATION” and “OBTENTION”, that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined. These axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope. Through the example of the IG molecular synthesis, the concepts generated from the “IDENTIFICATION”, “DESCRIPTION”, “CLASSIFICATION” and “NUMEROTATION” axioms are detailed with their main instances and semantic relations. The axioms have been essential for the conceptualization of the molecular immunogenetics knowledge and can be used to generate concepts for multi scale approaches at the molecule, cell, tissue, organ, organism or population level, emphasizing the generalization of the application domain. In that way the Formal IMGT-ONTOLOGY represents a paradigm for the elaboration of ontologies in system biology.

© 2007 Elsevier Masson SAS. All rights reserved.

Keywords: IMGT; Ontology; System biology; Immunogenetics; Immunoinformatics

1. Introduction

Immunogenetics, the science that studies the genetics of the immune responses, has shown a considerable expansion in biomedical fields since the last decades. It has highlighted the complex mechanisms by which B cells and T cells are at

the origin of the extreme diversity of antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different immunoglobulins and 10^{12} different T cell receptors per individual, in humans) [1,2]. These mechanisms include in particular DNA rearrangements [3] and, for the IG, somatic hypermutations [1,2]. In addition, there is a considerable polymorphism of the major histocompatibility complex (MHC), human leucocyte antigens (HLA) in humans. These particularities of the adaptive immune system of the vertebrates, and a better knowledge of the innate immune response found in any species, allow the immune system to be an excellent model for system biology. The huge amount of immunological experimental data

* Corresponding author. IMGT[®], the international ImMunoGeneTics information system[®], Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel.: +33 4 99 61 99 65; fax: +33 4 99 61 99 01.

E-mail address: marie-paule.lefranc@igh.cnrs.fr (M.-P. Lefranc).

continues to grow exponentially and necessitates to be managed and analysed computationally. This is the goal of immunoinformatics, a new emerging science that implements the bioinformatics methodologies to answer these needs. At the same time, standardized representation of genomics, genetics, proteomics and three-dimensional (3D) structures data is required to organize immunogenetics knowledge towards system biology and for the modelling and a better understanding of the immune system.

IMGT[®], the international ImMunoGeneTics information system[®] (<http://imgt.cines.fr>) [4], is the international reference in immunogenetics and immunoinformatics. Created in 1989

at the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) by Marie-Paule Lefranc (Université Montpellier 2 and CNRS) in Montpellier, France, IMGT provides a high-quality integrated knowledge resource, specialized in the IG, TR, MHC of human and other vertebrates, and related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF) of any species. The IMGT information system consists of databases (three of sequences, one of genes and one of 3D structures) and interactive tools for sequence, genome and 3D structure analysis, which interact together according to genomic, genetic and structural approaches [5] (Fig. 1).

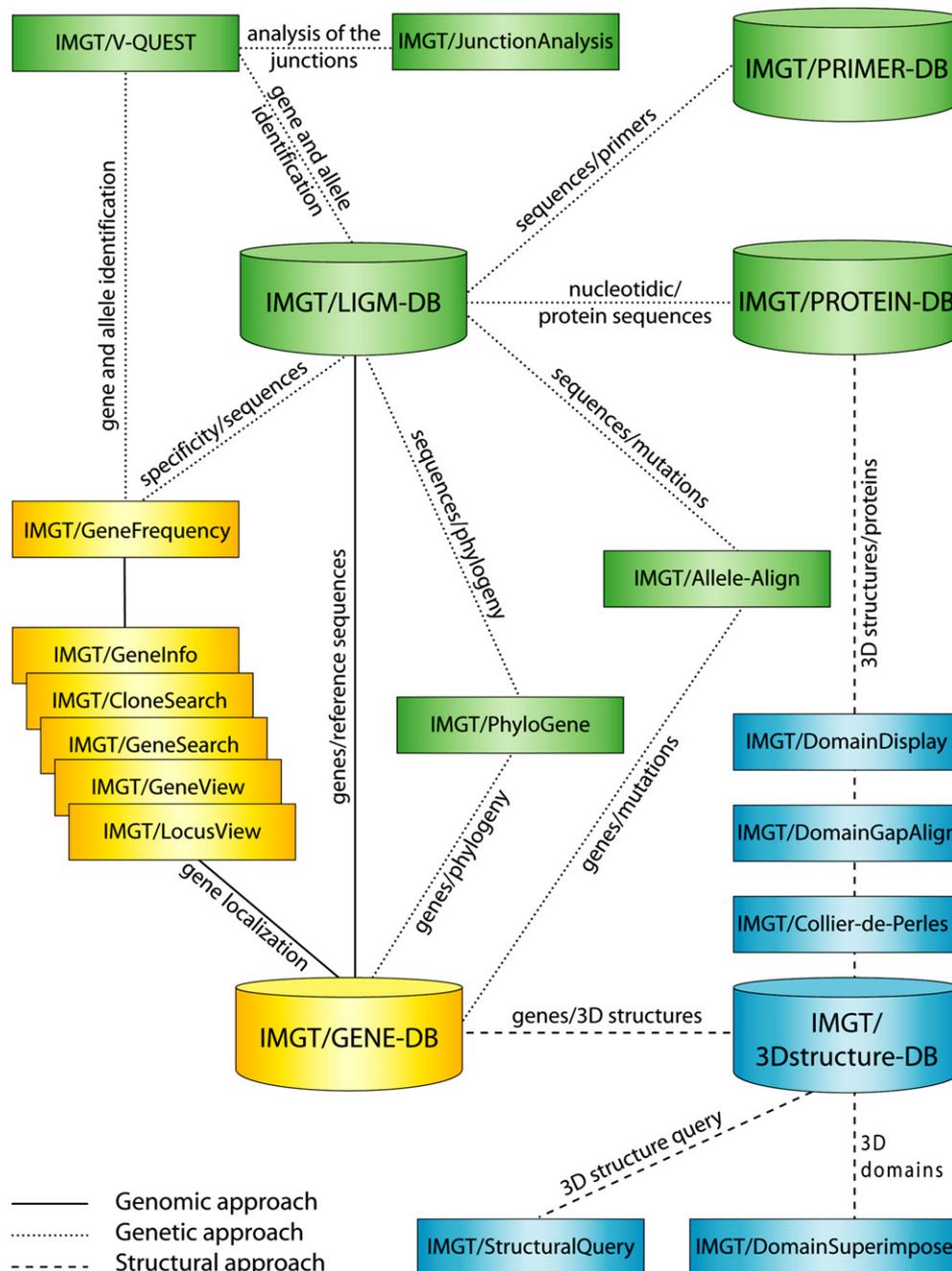


Fig. 1. IMGT[®] databases and tools with their interactions according to the genomic, genetic and/or structural approaches.

Moreover, IMGT provides Web resources comprising more than 10,000 HTML pages of synthesis (IMGT Repertoire), knowledge (IMGT Scientific chart, IMGT Education, IMGT Index) and external links (IMGT Bloc-notes and IMGT other accesses) [4].

The accuracy and the consistency of the IMGT data, as well as the coherence between the different IMGT components (databases, tools and Web resources), are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [6]. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in immunogenetics and immunoinformatics and manages the related knowledge, thus allowing the standardization for immunogenetics data from genome, proteome, genetics and 3D structures [7–9]. IMGT-ONTOLOGY results from a deep expertise in the domain and an extensive effort of conceptualization. The first standardization step was the identification of the IG and TR nucleotide sequences and the second step their description which led to the creation of IMGT/LIGM-DB [10], the first on-line IMGT database. The resulting controlled vocabulary comprises a thesaurus of keywords for the sequence identification and a set of labels for the description of the constitutive motifs. The third standardization step was the classification of the IG and TR genes which gave rise to the IMGT gene nomenclature for IG and TR of human and other vertebrates [1,2], approved by the Human Genome Organisation (HUGO) Nomenclature Committee HGNC in 1999 [11] and currently used in the generalist genome databases. The fourth standardization step was the setting up of the principles for the unique numbering of antigen receptor sequences and structures [12–16].

The standardization rules, defined in the IMGT Scientific chart [4], are based on the concepts of identification, description, classification and numerotation which characterize IMGT-ONTOLOGY [6] and which, interestingly, were defined before the term “ontology” became commonly used in biology and bioinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. Four of these axioms, “IDENTIFICATION”, “DESCRIPTION”, “CLASSIFICATION” and “NUMEROTATION” are presented in this paper, with the concepts that have been essential for the conceptualization of the molecular immunogenetics knowledge. As the same axioms can be used to generate concepts for multi-scale level approaches, the Formal IMGT-ONTOLOGY represents a paradigm for system biology ontologies, which need to identify, to describe, to classify and to numerotate objects, processes and relations at the molecule, cell, tissue, organ, organism or population levels.

2. Methods

2.1. Terminology

An ontology is a formal representation of a knowledge domain [6,17–19]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, “IDENTIFICATION”, “CLASSIFICATION”,

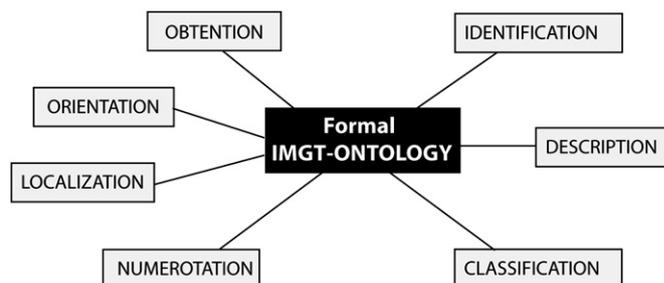


Fig. 2. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

“DESCRIPTION”, “LOCALIZATION”, “NUMEROTATION”, “ORIENTATION” and “OBTENTION”. These axioms postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined (Fig. 2). The axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope.

Each axiom gives rise to a set of concepts. Concepts are general in the reality [6,20–23]. Concept instances correspond to all possible examples of representation of a concept at a given granularity. A concept may be exemplified by one or several concept instances. New concept instances may be defined with the advancement of science. Concepts are linked by relations, the simplest being “is_a” which represents the edge between concepts at different levels of granularity and organizes the main hierarchy of IMGT-ONTOLOGY. Properties are semantic characteristics of a concept or of a concept instance: they may be simple attributes as a name alias, or they may be specific relations between concepts and instances across the main hierarchy. These relations are fundamental since they reveal strong semantic constraints and dependencies on which relies the coherence within or between IMGT components.

2.2. An example of knowledge at the molecular level: the immunoglobulin synthesis

The immunoglobulin synthesis, an example of knowledge at the molecular level, will be used to define the concepts generated by four of the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. The concepts of identification (IDENTIFICATION axiom) identify the nucleotide and protein sequences and the 3D structures according to a structured terminology, the concepts of description (DESCRIPTION axiom) describe the composition of the sequences and structures with standardized labels, the concepts of classification (CLASSIFICATION axiom) classify the genes and alleles with a standardized nomenclature, and the concepts of numerotation (NUMEROTATION axiom) numerotate the nucleotide and amino acid numbering within sequences and structures.

An IG or antibody is composed of two identical heavy chains associated with two identical light chains, kappa or lambda. In humans, heavy chain genes (locus IGH), light chain kappa genes (locus IGK) and light chain lambda genes (locus IGL) are located on the chromosomes 14 (14q32.3), 2 (2p11.2) and 22 (22q11.2), respectively. The synthesis of an immunoglobulin

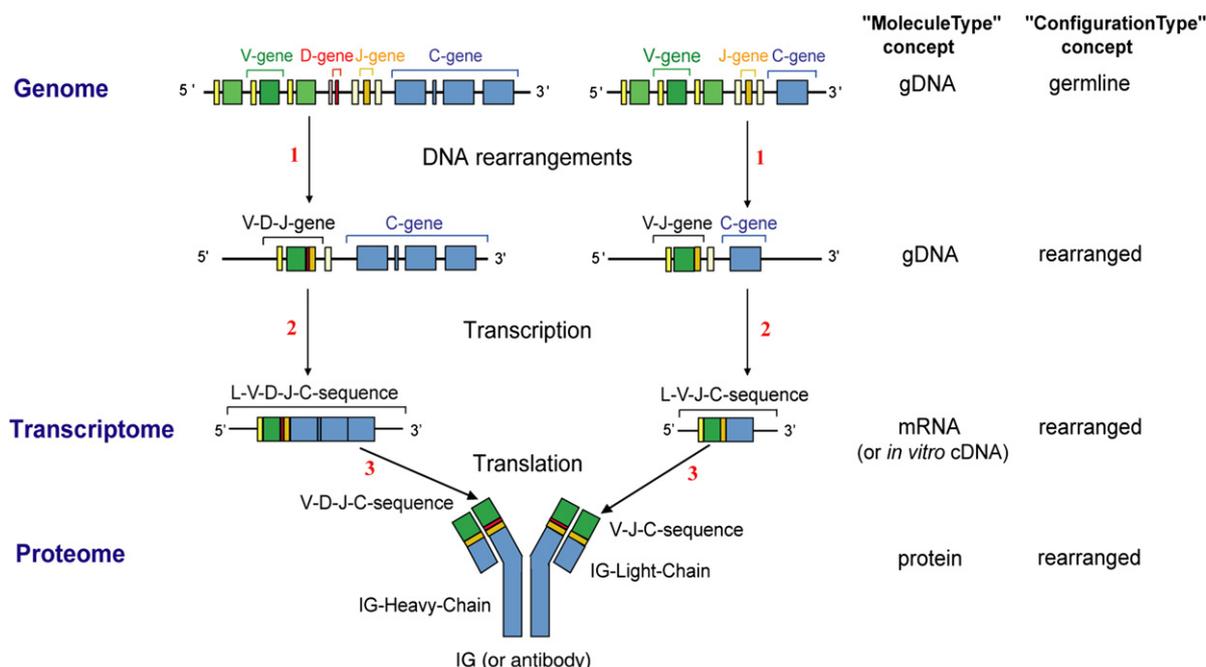


Fig. 3. An example of knowledge at the molecular level: the synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies [1]. 1: DNA rearrangements (is_rearranged_into), 2: Transcription (is_transcribed_into), 3: Translation (is_translated_into). The configuration of C-gene is undefined.

requires rearrangements of the IGH, IGK and IGL genes during the differentiation of the B lymphocytes.

In the human genome (genomic DNA or gDNA), four types of genes code the IG (and TR): variable (V), diversity (D), joining (J) and constant (C). The configuration of the V-gene, D-gene and J-gene is identified as “germline” (Fig. 3), the configuration of the C-gene is “undefined”. During the differentiation of the B lymphocytes in the bone marrow, the genomic DNA is rearranged first in the IGH locus, and then in the IGK and IGL loci. The rearrangements in the IGH locus lead to the junction of a D-gene and a J-gene to form a D-J-gene, and then to the junction of a V-gene to the D-J-gene to form a V-D-J-gene. The rearrangements in the IGK or IGL loci lead to the junction of a V-gene and a J-gene to form a V-J-gene. The configuration of these genes is identified as “rearranged”. After transcription and maturation of the pre-messenger by splicing, the messenger RNA (mRNA) L-V-D-J-C-sequence and L-V-J-C-sequence (L for leader) are obtained and then translated into the heavy chain (IG-Heavy-Chain) and the light chain (IG-Light-Chain) of an IG (or antibody) (Fig. 3).

The variable domains VH and VL are coded by the V-D-J-REGION and the V-J-REGION (Fig. 4). Each domain includes four framework regions (FR) (in pale grey in Fig. 4) and three hypervariable loops or complementarity determining regions (CDR). The CDR, and more particularly the CDR3 that result from the junction of the V-D-J genes (in the VH domain) and V-J genes (in the VL domain), are involved in the antigen recognition. The VH and VL amino acids in contact with the antigen constitute the paratope. The part of the antigen recognized by the antibody is the epitope. The number of

potential V-D-J and V-J rearrangements depends on the number of functional V, D and J genes in the genome. Additional mechanisms (N diversity at the V-D-J and V-J junctions and somatic hypermutations) allow to reach 10^{12} different antibodies per individual [1] (IMGT®, <http://imgt.cines.fr>).

2.3. Implementation

The main hierarchy of the IMGT-ONTOLOGY concepts has previously been described [6]. IMGT-ONTOLOGY

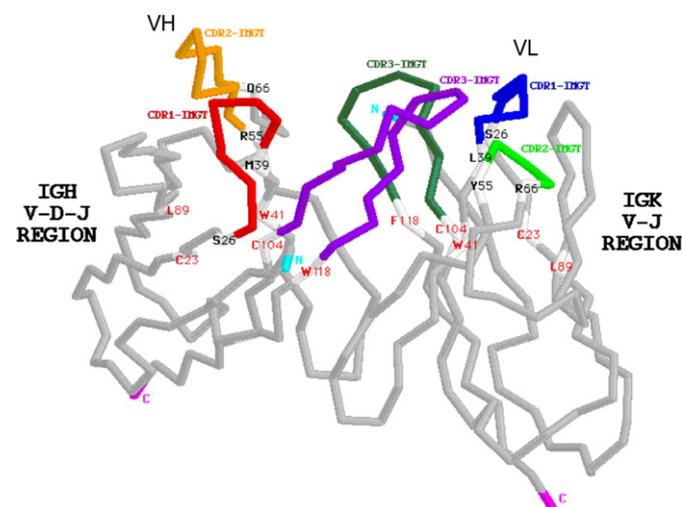


Fig. 4. The variable domains VH and VL of the heavy and light chains of an IG or antibody. VH CDR1-IMGT is in red, CDR2-IMGT in orange and CDR3-IMGT in purple. VL CDR1-IMGT is in blue, CDR2-IMGT in green and CDR3-IMGT in dark green.

concepts are available for the biologists and IMGT users in natural language in the IMGT Scientific chart [4], and have been formalized for programming purpose in IMGT-ML [24,25] which is an XML Schema (<http://www.w3.org/TR/xmlschema-0/>). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé [26] and OBO-Edit (<http://oboedit.org/>), that are frequently used ontology editors for biological ontologies. Protégé and OBO-Edit ontologies can be exported into RDF (<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>) and OWL (<http://www.w3.org/2004/OWL/>) which allow interoperability with other ontologies.

3. Results

3.1. The necessity of identification: the IDENTIFICATION axiom

The IDENTIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and relations, have to be identified. The IDENTIFICATION axiom has generated the concepts of identification which provide the terms and rules to identify an entity, its processes and its relations. In molecular biology, the concepts of identification allow to identify the molecules, their processes and their relations at the genome, transcriptome and proteome levels.

3.1.1. Identification of an organism: the “Taxon” concept

The “Taxon” concept allows to identify the type of taxon in which an object, process or relation is found. The “Taxon” concept manages a hierarchy of concepts at various levels of granularity. The corresponding hierarchical taxonomy is that provided by the National Center for Biotechnology Information NCBI (<http://www.ncbi.nlm.nih.gov/>) up to the rank of species (“Species” concept) and subspecies (“Subspecies” concept) in order to establish complete interoperability with generalist databases. Since IG, TR and MHC genes are only present in jawed vertebrates (gnathostoma), only vertebrate species were originally represented in IMGT-ONTOLOGY. However, with the extension of IMGT-ONTOLOGY to the IgSF and MhcSF, invertebrate species are incorporated whenever necessary. The “EthnicGroup”, “Breed” and “Strain” concepts have been added to IMGT-ONTOLOGY to allow the identification of data specific to ethnic groups for humans (http://www.ebi.ac.uk/imgt/hla/help/ethnic_help.html), breeds for domestic animals or strains for laboratory [27] and wild animals.

3.1.2. Identification of an entity: the “EntityType” concept

The “EntityType” concept identifies the type of entity. An entity can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the “EntityType” concept is designated as “Molecule_EntityType”, which is defined by the “MoleculeType”, “GeneType” and “ConfigurationType” concepts of identification and has properties

identified in the “Functionality” and “StructureType” concepts (Fig. 5).

3.1.2.1. The “MoleculeType” concept. The “MoleculeType”, concept identifies the type of molecule based on the type of the constitutive elements and on the concepts of obtention (not detailed here). The four main instances of the “MoleculeType” concept are ‘gDNA’ (genomic DNA, a nucleotide sequence made of A, T, C, G, obtained from a genome), ‘mRNA’ (messenger RNA or transcript, a nucleotide sequence made of A, U, C, G, obtained by transcription of a genomic DNA), ‘cDNA’ (complementary DNA, a nucleotide sequence made of A, T, C, G, obtained *in vitro* by reverse transcription of the messenger RNA) and ‘protein’ (a sequence made of amino acids, obtained by translation of a transcript). Thus, the instances of the “MoleculeType” concept allow to identify a sequence: nucleotide sequence that can be either genomic (‘gDNA’) or a transcript (‘mRNA’, ‘cDNA’), and amino acid sequence (‘protein’).

3.1.2.2. The “GeneType” concept. The “GeneType” concept identifies the type of gene and comprises five instances (Fig. 5). The first instance, ‘conventional’, refers to any (coding or not coding) gene other than IG or TR genes. The other four instances are specific to immunogenetics: ‘variable’ (V), ‘diversity’ (D) and ‘joining’ (J) gene types that rearrange at the DNA level and code the variable domains of IG and TR, and ‘constant’ (C) gene type that codes the constant region of IG and TR [1,2].

3.1.2.3. The “ConfigurationType” concept. The “ConfigurationType” concept identifies the type of gene configuration and comprises three instances (Fig. 5). The instance ‘undefined’ identifies the configuration of the conventional and of the constant (C) genes. The instances ‘germline’ and ‘rearranged’ identify the status of the V, D and J genes, before and after DNA rearrangements, respectively [1,2].

3.1.2.4. The “Molecule_EntityType” concept. The “Molecule_EntityType” concept, defined by the “MoleculeType”, “GeneType” and “ConfigurationType” concepts, includes 19 instances. Three instances, ‘gene’, ‘nt-sequence’ and ‘AA-sequence’, respectively identify the gDNA, mRNA and protein (“MoleculeType”) of a conventional gene (“GeneType”) in undefined configuration (“ConfigurationType”). The nt-sequence instance is also valid for cDNA. Sixteen instances allow to identify the IG and TR. Ten of them are represented in Fig. 3: six for the gDNA (‘V-gene’, ‘D-gene’, ‘J-gene’, ‘C-gene’, ‘V-D-J-gene’ and ‘V-J-gene’), two for the mRNA, ‘L-V-D-J-C-sequence’ and ‘L-V-J-C-sequence’, also valid for cDNA, and two for the protein, ‘V-D-J-C-sequence’ and ‘V-J-C-sequence’. For example, the instance ‘V-gene’ identifies a gDNA (“MoleculeType”) containing a gene V (“GeneType”), in germline configuration (“ConfigurationType”). The instance ‘L-V-J-C-sequence’ identifies a sequence of mRNA or cDNA (“MoleculeType”) corresponding to V, J and C genes (“GeneType”), in rearranged configuration

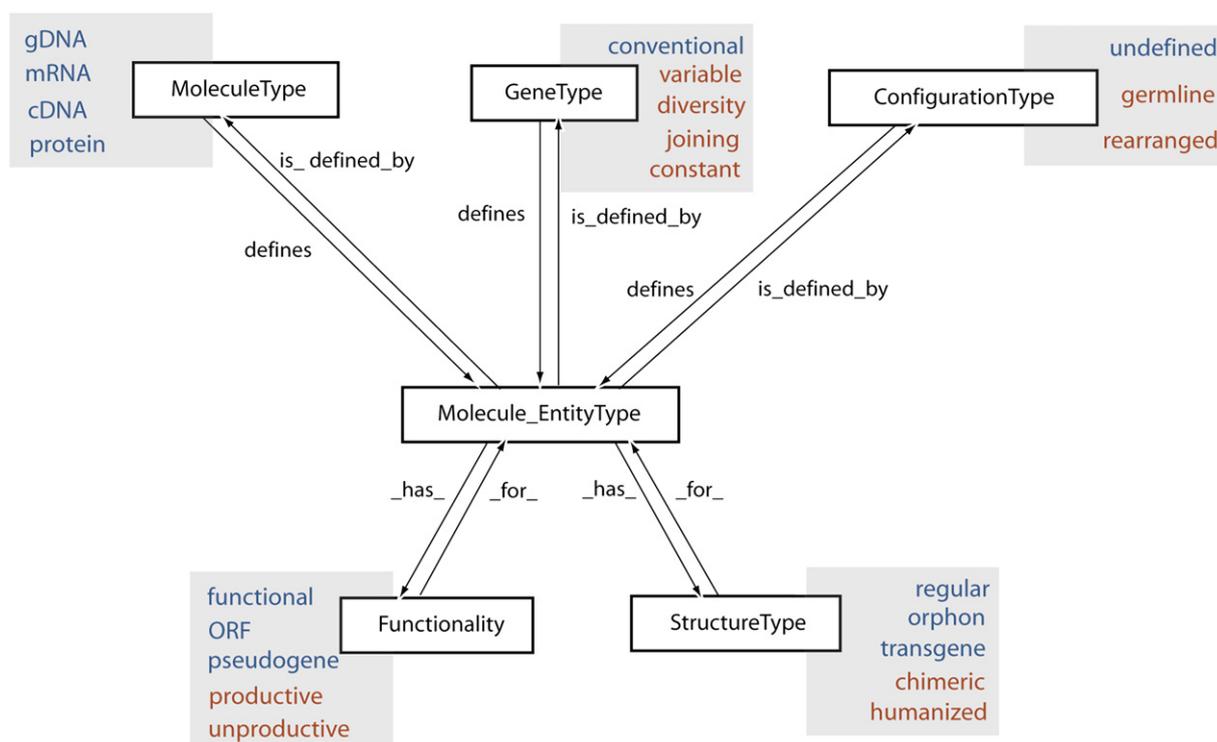


Fig. 5. The “Molecule_EntityType” concept. The “Molecule_EntityType” concept is defined by the “MoleculeType”, “GeneType” and “ConfigurationType” concepts of identification and has properties identified in the “Functionality” and “StructureType” concepts (IDENTIFICATION axiom). Arrows indicate reciprocal relations “is_defined_by” and “defines”, “_has_” and “_for_”. Concept instances which are general are in blue, those which are specific of the IG and TR are in red. The “Molecule_EntityType” concept has 19 instances (listed in Section 3.1.2.4). Only a few examples of the “StructureType” concept instances are shown.

(“ConfigurationType”) (Fig. 3). The last six instances correspond to partial rearrangement (‘D-J-gene’) or to sterile transcripts (‘L-V-sequence’, ‘D-sequence’, ‘J-sequence’, ‘J-C-sequence’ and ‘C-sequence’).

3.1.2.5. The “Functionality” concept. The “Functionality” concept identifies the type of functionality for the “Molecule_EntityType” concept (Fig. 5). It includes five instances, divided into two categories, according to the configuration type. Three instances, ‘functional’, ‘ORF’ (open reading frame) and ‘pseudogene’ identify the functionality of a “Molecule_EntityType” instance in undefined or germline configuration. They allow to identify the functionality of conventional genes, that of C genes, and that of V, D and J genes before their rearrangement in the genome, and by extension the functionality of their transcripts and proteins. The two instances ‘productive’ and ‘unproductive’ identify the functionality of “Molecule_EntityType” instances in rearranged configuration. They allow to identify the functionality of IG and TR entities after their rearrangement in the genome, that of fusion genes resulting from translocations, and that of hybrid genes obtained by biotechnology molecular engineering, and by extension the functionality of their transcripts and proteins.

3.1.2.6. The “StructureType” concept. The “StructureType” concept identifies the structure for the “Molecule_EntityType” concept. This concept allows to identify structures

with a classical organization (‘regular’), from those which have been modified either naturally *in vivo* (‘orphon’, ‘processed orphon’, ‘unprocessed orphon’, ‘unspliced’, ‘partially spliced’, etc.), or artificially *in vitro* (‘chimeric’, ‘humanized’, transgene, etc.).

3.1.3. Identification of a receptor: the “ReceptorType” concept

The “ReceptorType” concept identifies the type of receptor. A receptor can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the “ReceptorType” concept is designated as “Molecule_ReceptorType” which is defined by the “ChainType” concept of identification and has properties identified in the “StructureType”, “Specificity” and “Function” concepts (Fig. 6). The “ChainType” concept is itself defined by the “Molecule_EntityType” and the “DomainType” concepts of identification and by concepts of classification (see CLASSIFICATION axiom). These latter are organized in a hierarchy which confers different levels of granularity to the “Molecule_ReceptorType” and “ChainType” concepts.

3.1.3.1. The “Molecule_ReceptorType” concept. The “Molecule_ReceptorType” concept identifies the type of protein receptor, defined by its chain composition. Thus, IG is an instance of the “Molecule_ReceptorType” concept, defined as comprising 4 chains, two heavy chains and two light chains,

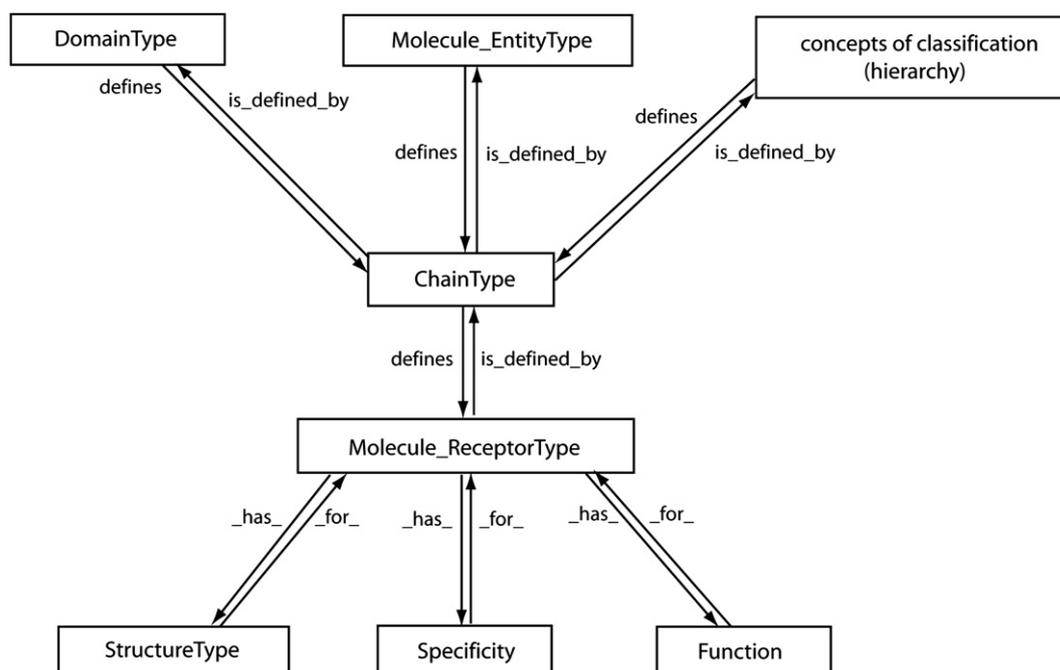


Fig. 6. The “Molecule_ReceptorType” concept. The “Molecule_ReceptorType” concept, defined by the “ChainType” concept of identification, has properties identified in the “StructureType”, “Specificity” and “Function” concepts (IDENTIFICATION axiom). The “ChainType” concept is itself defined by the “Molecule_EntityType” and “DomainType” concepts and by concepts of classification (hierarchy). Arrows indicate reciprocal relations “is_defined_by” and “defines”, “_has_” and “_for_”. These concepts have different levels of granularity, up to six for “Molecule_ReceptorType” and “ChainType”.

identical two by two and covalently linked (Fig. 7). A receptor may comprise one chain (monomer) or several associated chains (multimer).

3.1.3.2. The “ChainType” concept. The “ChainType” concept identifies the type of chain (Fig. 6). It is one of the most important concepts of identification for the standardization of genome, transcriptome and proteome data in system biology. Indeed, being able to identify a type of chain means that it is possible to identify the transcript and the encoding gene(s). The “ChainType” concept contains a hierarchy of concepts which identify the chain type at different levels of granularity. The finest level of granularity, the “GeneLevel-ChainType” concept, identifies the type of chain by reference to the gene(s) which code(s) the chain. It represents the main concept for a very precise identification because it establishes a relationship with the “Gene” concept which belongs to the concepts of classification (reciprocal relations “is_coded_by” and “codes”). The number of instances of the “GeneLevel-ChainType” concept depends on the number of functional genes and ORF per haploid genome in a given species (in the case of the IG and TR, it is the number of functional and ORF constant genes which is taken into account). If only the functional genes are considered, the instances of this concept correspond to the isotypes.

3.1.3.3. The “DomainType” concept. A chain can be defined by its constitutive structural units (“DomainType” concept) (Fig. 6). A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino

acid sequence and the nucleotide sequence which encodes it. This concept may theoretically comprise many instances, but so far only the instances which have been carefully characterized by LIGM have been entered in IMGT-ONTOLOGY. The “DomainType” concept has currently three instances, V type domain (variable domains of the IG and TR and V-like domains of other IgSF proteins), C type domain (constant domains of the IG and TR and C-like domains of other IgSF proteins) and G type domain (groove domains of the MHC and G-like domains of other MhcSF proteins) [14–16].

3.1.3.4. The “Specificity” and “Function” concepts. The “Specificity” concept identifies the specificity of the “Molecule_ReceptorType” (Fig. 6), and by extension the specificity of the chains and domains and of the corresponding transcripts. Instances of the “Specificity” concept identify the antigen recognized by an antigen receptor (IG or TR). The “Specificity” concept is particularly important because of the unlimited number of antigens and of the complexity of the antigen/antigen receptor interactions. The conceptualization of knowledge associated with this concept is in the course of modelling. The instances of the “Specificity” concept (several hundreds at the present time) will be connected on the one hand, with the “Epitope” concept which identifies the part of the antigen recognized by the antigen receptor and on the other hand, with the “Paratope” concept which identifies the part of the antigen receptor (IG or TR) which recognizes and binds to the antigen. The “Function” concept identifies the function of the “Molecule_ReceptorType” (Fig. 6), and by extension the function of the chains and domains and of the corresponding

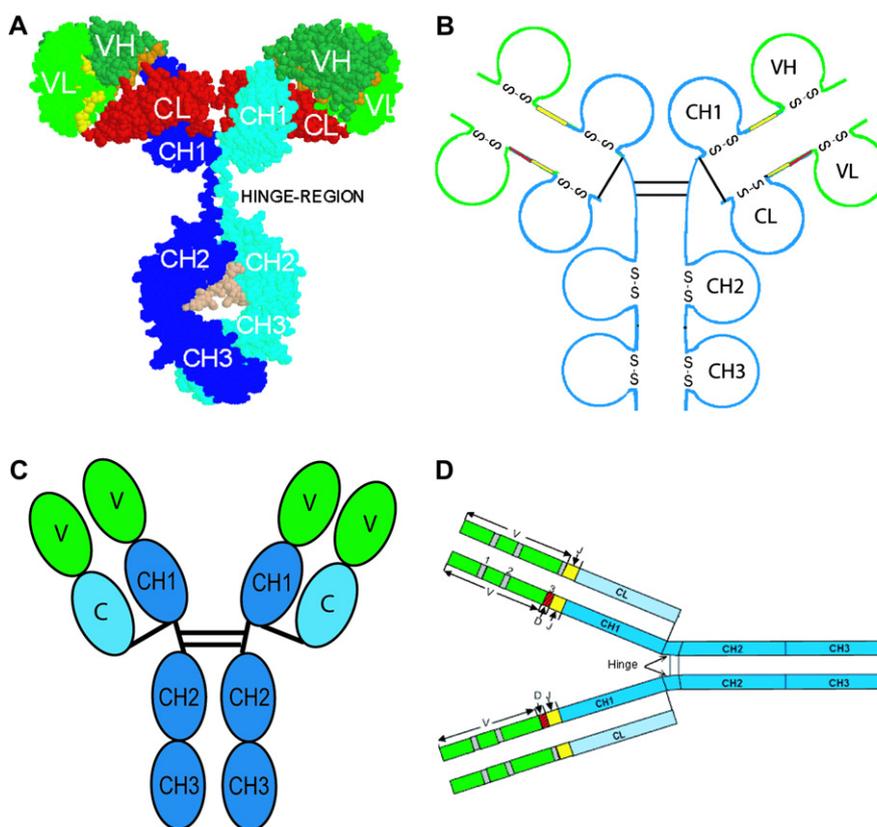


Fig. 7. Identification of an IG or antibody as an instance of the “Molecule_ReceptorType” concept made of four chains, two IG-Heavy-Chain and two IG-Light-Chain (“ChainType” concept). The four representations, although different, allow to identify an IG as a receptor of four chains, themselves organized in domains (“DomainType” concept). VH and VL are V type domains, coded by the V-D-J region and V-J region, respectively. CL, CH1, CH2 and CH3 are C type domains. (A) 3D structure, (B) organization in Ig-like domains, (C) organization in modules, (D) regions coded by the V, D, J and C gene types. The C gene type codes the constant region (CL for the IG-Light-Chain and CH1, hinge, CH2 and CH3 for the IG-Heavy-Chain). This representation, schematized as a Y shape, is frequently used to represent an IG.

transcripts. Instances of the “Function” concept identify the dual function of the antigen receptors [2]. Their identification and definition are still in development.

3.2. The necessity of description: the DESCRIPTION axiom

The DESCRIPTION axiom of the Formal IMGTOLOGY or IMGTKaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be described.

3.2.1. Description of an entity: the “EntityPrototype” concept

The “EntityPrototype” concept, generated from the DESCRIPTION axiom, provides the description of the “EntityType” concept (IDENTIFICATION axiom). Each instance of the “EntityPrototype” concept is linked to an instance of the “EntityType” concept by the reciprocal relations “describes” and “is_described_by”. The “EntityPrototype” concept allows the description of the entity organization and of its constitutive motifs. The “Core” concept allows to describe the parts of the entities which need to be described in all instances

of the “EntityPrototype” concept. These two concepts of description, “Molecule_EntityPrototype” and “Core”, which have been particularly highlighted by IMGTO, are described below as examples.

3.2.2. The “Molecule_EntityPrototype” concept

In molecular biology, the DESCRIPTION axiom has generated the concepts of description which provide the terms and the rules to describe motifs in the nucleotide and protein sequences and in 3D structures. These concepts gave rise to a standardized terminology and to a precise definition of the annotation rules. The ontology for sequences and 3D structures has been the focus of IMGTO for many years. The instances of the concepts of description correspond to IMGTO labels. More than 550 labels were defined (270 for the nucleotide sequences (<http://imgt.cines.fr/cgi-bin/IMGTOlect.jv?query=7>) [10] and 285 for the 3D structures [28] (<http://imgt.cines.fr/textes/IMGTOScientificChart/SequenceDescription/IMGTO3Dlabeldef.html>). Interestingly, 64 IMGTO labels defined for nucleotide sequences are used and cross-referenced in the recently created Sequence Ontology (SO) (<http://song.sourceforge.net/>) [29] to describe specific IG and TR gene organization (<http://imgt.cines.fr/textes/IMGTOindex/ontology.html>).

The “Molecule_EntityPrototype” concept allows the description of the entity (gene, transcript and protein) organization and of their constitutive motifs. This concept is fundamental in IMGT-ONTOLOGY because it allows the representation of the knowledge related to the complex mechanisms of IG and TR gene rearrangements (Fig. 8). The relation “is_rearranged_into” is specific to the synthesis of the IG and TR. The relations “is_transcribed_into” and “is_translated_into” are general for molecular biology. These three relations allow the organization of the various instances of the “Molecule_EntityPrototype” concept during the synthesis of the IG and the TR, and in a more general way for the expression of any protein. They allow in addition, by more specific relations, to take into account the alternative transcripts, the protein isoforms and the post-translational modifications.

Each of the 19 instances of the concept “Molecule_EntityPrototype” can be described with its constitutive motifs which belong to the other concepts of description. Thus Fig. 9 shows as examples the graphical representation of the V-GENE and V-D-J-GENE instances with their constitutive motifs.

A set of ten relations are necessary and sufficient to compare the localization of the motifs of an instance of the concept “Molecule_EntityPrototype” (Table 1). These relations are part of the concepts of localization (LOCALIZATION axiom) (IMGT Index, <http://imgt.cines.fr>).

3.2.3. The “Core” concept

The “Core” concept allows to describe the coding region of genes and contains five instances which are ‘REGION’ (for conventional gene type), ‘V-REGION’, ‘D-REGION’, ‘J-REGION’ and ‘C-REGION’ (for V, D, J and C gene types, respectively). These instances are particularly important since they can be described in all the instances of the “Molecule_EntityPrototype”

concept. They allow to describe the chains of the antigen receptors in spite of the complexity of their structure and to link sequences, structures and functions. Moreover, these are the instances of the “Core” concept which allowed the definition and standardized description of the IG and TR alleles (concepts of classification), now approved at the international level [1,2].

3.3. The necessity of classification: the CLASSIFICATION axiom

The CLASSIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be classified. In molecular biology, the concepts of classification generated from the CLASSIFICATION axiom allow to classify and name the genes and their alleles. The genes which code the IG and TR belong to highly polymorphic multigenic families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature [1,2] (Fig. 10). The IMGT gene nomenclature has been approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC), in 1999 [11]. The IMGT IG and TR gene names are the official reference for the genome projects and, as such, have been integrated in the Genome Database (GDB), in LocusLink and in Entrez Gene at NCBI [30]. The IG and TR genes [1,2] are managed in the IMGT/GENE-DB database [31].

3.3.1. The “Group” and “Subgroup” concepts

The “Group” concept classifies a set of genes which belong to the same multigene family, within the same species

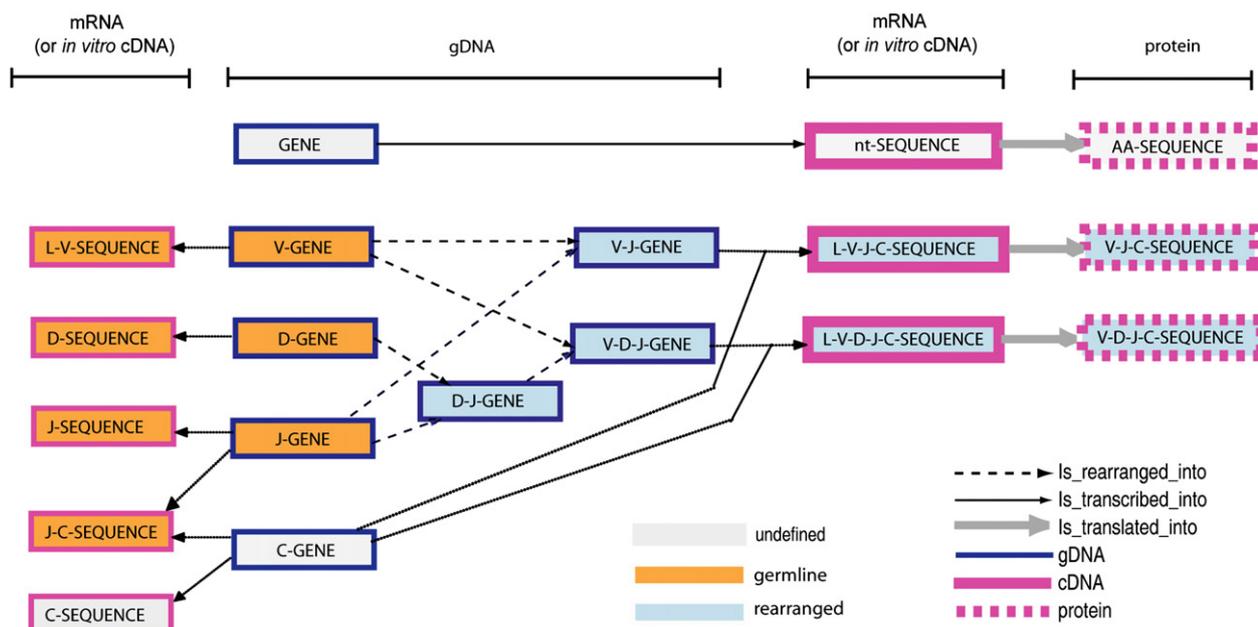


Fig. 8. Instances of the “Molecule_EntityPrototype” concept (DESCRIPTION axiom). The three instances “GENE”, “nt-SEQUENCE” and “AA-SEQUENCE” correspond to conventional genes while the 16 other instances are specific of the IG and TR. The concept instances for mRNA are also valid for *in vitro* cDNA. The first column corresponds to ‘sterile transcript’ instances.

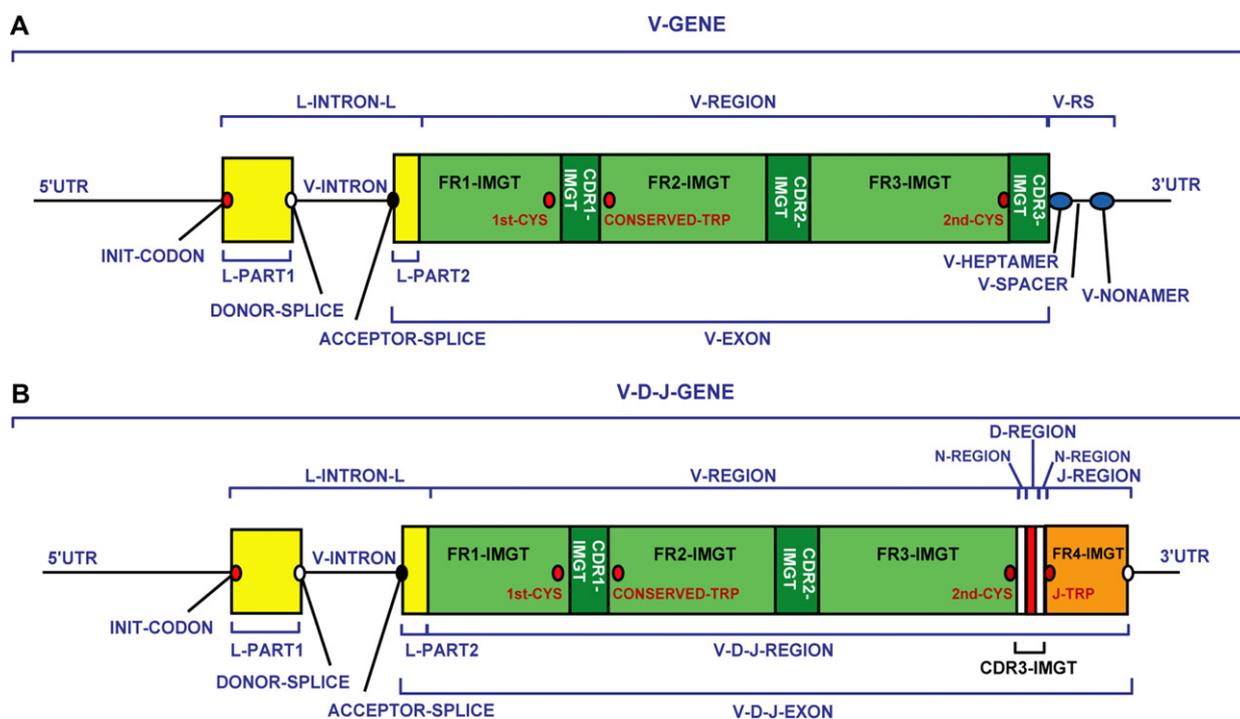


Fig. 9. Graphical representation of two instances of the “Molecule_EntityPrototype” concept (DESCRIPTION axiom). (A) V-GENE. (B) V-D-J-GENE. Twenty-five labels and ten relations are necessary and sufficient for a complete description of these instances.

or between different species. For the IG and TR, the set of genes is identified by an instance of the “GeneType” concept (V, D, J or C). The “Subgroup” concept classifies a subset of genes which belong to the same group, and which, in a given species, share at least 75% of identity at the nucleotide sequence level (and in the germline configuration for the V, D, and J genes).

3.3.2. The “Gene” and “Allele” concepts

The “Gene” concept classifies a unit of DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5' and 3', and the introns, if present). The instances of the “Gene” concept are gene names. In IMGT-ONTOLOGY, a gene name is composed of the name of the species (instance of the Taxon “Species” concept) and of the international HGNC/IMGT gene symbol, for example, *Homo sapiens* IGLV1–2. By extension, orphans and pseudogenes are also instances of the “Gene” concept. The “Allele” concept classifies a polymorphic variant of a gene. The instances of the “Allele” concept are allele names. Alleles identified by the mutations of the nucleotide sequence are classified by reference to allele *01.

Table 1
Relations for sequence description (LOCALIZATION axiom)

Relation	Reciprocal relation
“adjacent_at_its_5_prime_to”	“adjacent_at_its_3_prime_to”
“included_with_same_5_prime_in”	“includes_with_same_5_prime”
“included_with_same_3_prime_in”	“includes_with_same_3_prime”
“overlaps_at_its_5_prime_with”	“overlaps_at_its_3_prime_with”
“included_in”	“includes”

Full description of mutations and allele name designations are currently recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). They are reported in Alignment tables, in IMGT Repertoire <http://imgt.cines.fr> and in IMGT/GENE-DB [16].

3.4. The necessity of numbering: the NUMEROTATION axiom

The NUMEROTATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be numerotated. So far, these concepts have essentially been defined at the molecular level. The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for a domain (sequences and 3D structures) [14–16] (Fig. 11). The “IMGT_unique_numbering” concept has three concept instances: “IMGT_unique_numbering_for_V_Type_domain”, “IMGT_unique_numbering_for_C_type_domain”, “IMGT_unique_numbering_for_G_type_domain” [14–16].

The “IMGT_unique_numbering” concept determines the “FR-IMGT_length”, “CDR-IMGT_length”, “Strand_length”, and “Helix_length” concepts [14–16]. The “IMGT_unique_numbering” concept is illustrated by the “IMGT_Collier_de_Perles” concept which allows graphical representation in two dimensions (2D) of the amino acid sequences of V, C or G type domains [32,33] and comprises three concept instances (Fig. 12). This concept is largely recognized at the international level and the expression “IMGT Collier de Perles” is now used in scientific publications.

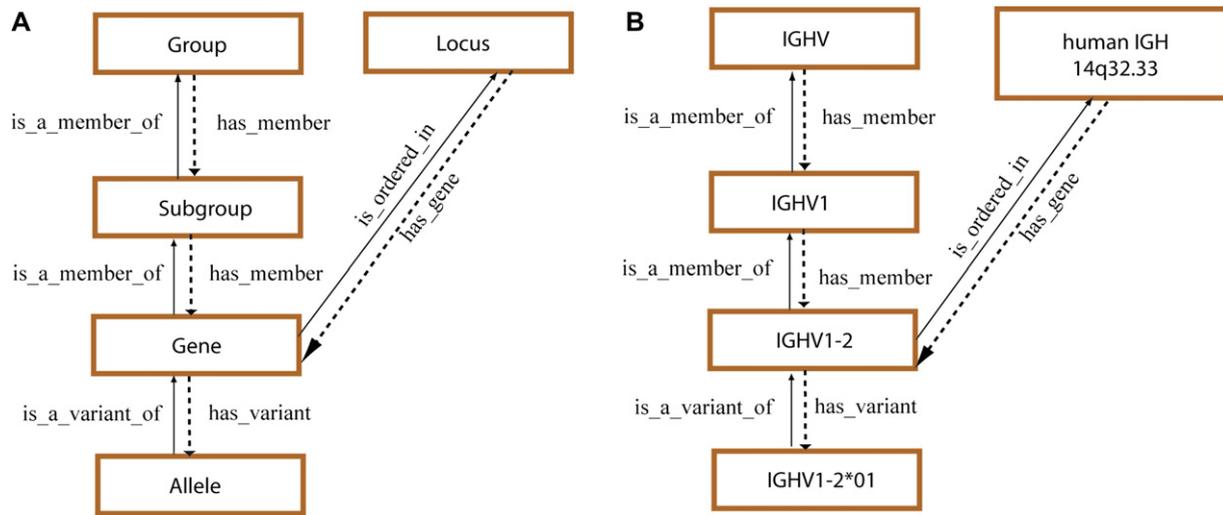


Fig. 10. Concepts of classification for gene and allele nomenclature (CLASSIFICATION axiom). (A) Hierarchy of the concepts of classification and their relations. (B) Examples of concept instances for each concept of classification. The concepts instances are associated to an instance of the “Taxon” concept, and more precisely for the “Gene” and “Allele” concepts to an instance of the “Species” concept (here, *Homo sapiens*). The “Locus” concept is a concept of localization (LOCALIZATION axiom).

The “IMGT_Collier_de_Perles” concept is particularly used in antibody engineering for the humanization of murine antibodies in which it is necessary to precisely delimit the murine CDR-IMGT to be grafted, in order to preserve the antibody specificity. The concepts of numerotation are also at the origin of the standardization of the allele description and, more generally of the mutation description (IMGT Scientific chart, <http://imgt.cines.fr>).

4. Conclusion

The inherent difficulties due to the complexity and diversity of immunogenetics knowledge gave rise to a conceptualization in IMGT-ONTOLOGY which has been developed on an original and unprecedented approach. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulate that the approach to manage biological data and to represent

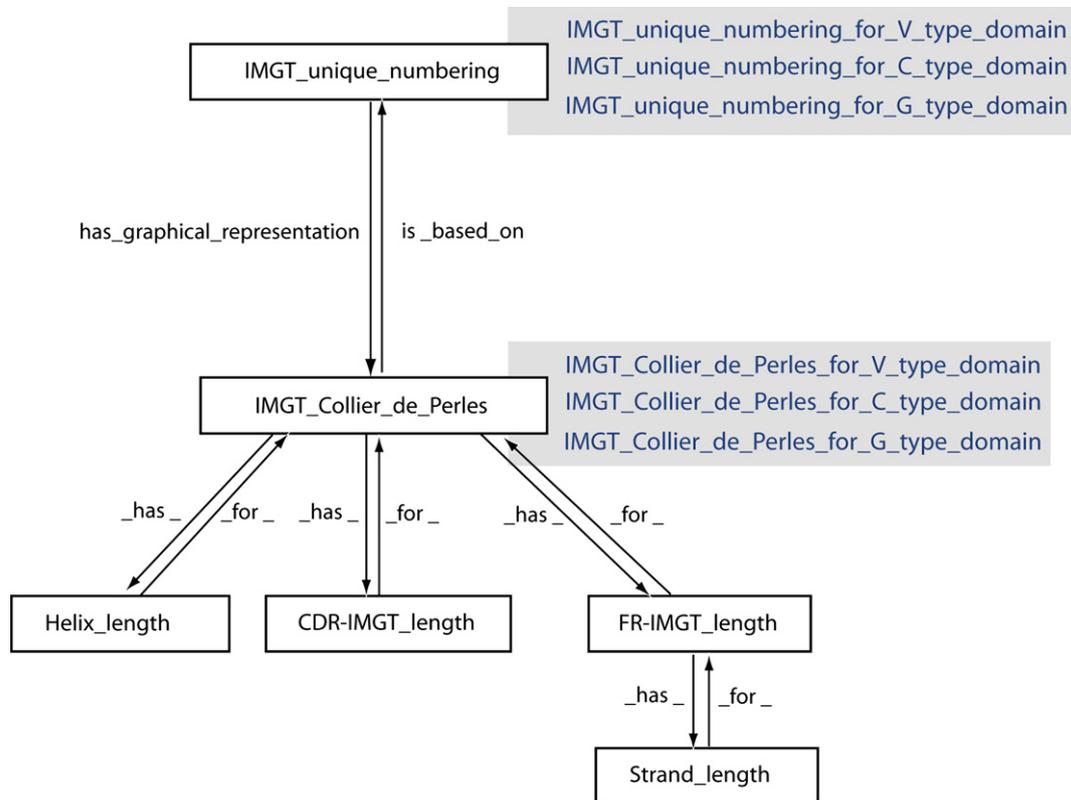
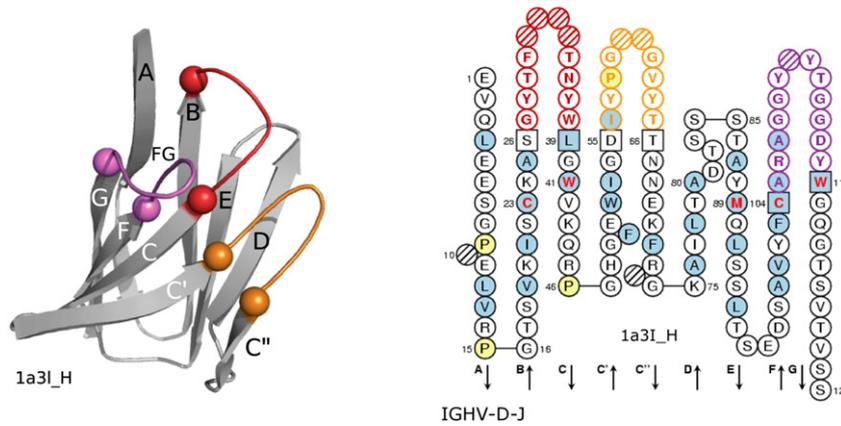


Fig. 11. The “IMGT_unique_numbering” and “IMGT_Collier_de_Perles” concepts and relations with other concepts of numerotation (NUMEROTATION Axiom). Concept instances are written in blue. Arrows indicate reciprocal relations “has_graphical_representation” and “is_based_on”, “_has_” and “_for_”.

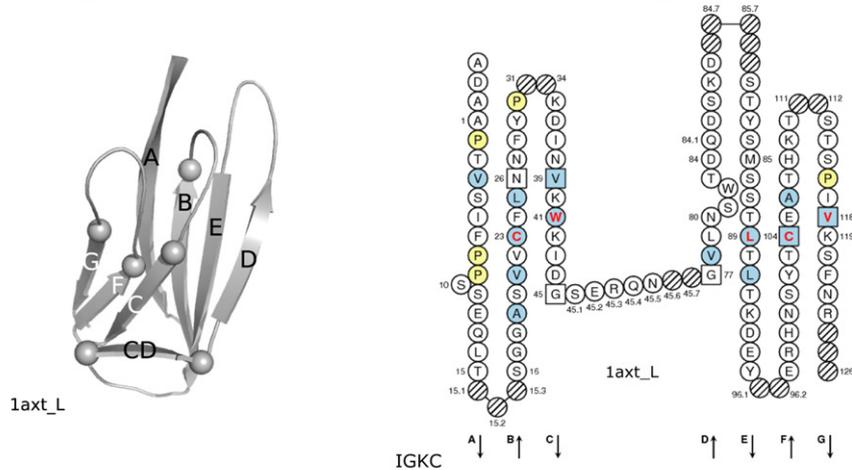
knowledge in biology comprises various facets. The IMGT-ONTOLOGY concepts generated from these axioms have allowed the representation, at the molecular level, of knowledge related to the genome, transcriptome, proteome, genetics and 3D structures. This multi-faceted approach has great potential for multi-scale system biology. Indeed, the IDENTIFICATION, DESCRIPTION, CLASSIFICATION and NUMEROTATION axioms are valid, not only for molecules, but also for cells, tissues, organs, organisms or populations. In addition, the

LOCALIZATION, ORIENTATION and OBTENTION axioms (in development) will allow the integration of the time and space concepts and the follow-up of the components and their changes of states and properties, as well as the definition and characterization of processes, functions and activities. Thus, IMGT-ONTOLOGY represents, by its 7 axioms and the concepts generated from them, a paradigm for the elaboration of ontologies in system biology which requires to identify, to describe, to classify, to numerotate, to localize, to orientate and to determine

A V type domain and IMGT_Collier_de_Perles_for_V_type_domain



B C type domain and IMGT_Collier_de_Perles_for_C_type_domain



C G type domain and IMGT_Collier_de_Perles_for_G_type_domain

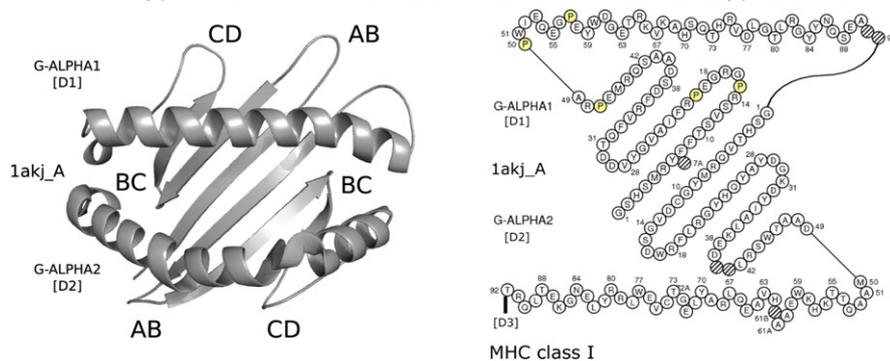


Fig. 12. “DomainType” and “IMGT_Collier_de_Perles” concept instances.

the obtaining and evolution of biological knowledge from molecule to population, in time and space.

The concepts of IMGT-ONTOLOGY are available, for the users of IMGT and the biologists in general, in natural language in IMGT Scientific chart (<http://imgt.cines.fr>), and have been formalized for programming purpose in IMGT-ML (XML Schema). IMGT-ONTOLOGY is being implemented in Protégé and OBO-Edit to facilitate the export in formats such as OWL, and to link, whenever possible, the concepts of IMGT-ONTOLOGY to those of other ontologies in biology such as the Gene Ontology (GO) [34], and in immunology, such as the Immunome Epitope database and Analysis Resource (IEDB) [35] and other Open Biomedical Ontologies (OBO) (<http://obo.sourceforge.net>).

The concepts of IMGT-ONTOLOGY are currently used for the exchange and the sharing of knowledge in very diverse fields of research at the molecular level: (i) fundamental and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukaemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wild life species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (scFv, phage displays, combinatorial libraries, chimaeric, humanized and human antibodies), (vi) diagnostics (clonalities, detection and follow-up of residual diseases) and (vii) therapeutic approaches (grafts, immunotherapy, vaccinology). IMGT-ONTOLOGY represents a key component in the elaboration and setting up of standards of the European ImmunoGrid project (<http://www.immunogrid.org/>) whose aim is to define the essential concepts for modelling of the immune system.

Acknowledgements

We are grateful to Gérard Lefranc for helpful discussion and to the IMGT[®] team for its constant motivation and its expertise. IMGT[®] was funded by the Centre National de la Recherche Scientifique (CNRS), the BIOMED1 (BIOCT 930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programmes of the European Union. IMGT received subventions from Association pour la Recherche sur le Cancer (ARC), Génopole-Montpellier-Languedoc-Roussillon and the Réseau National des Génopoles (RNG). IMGT has been a National Bioinformatics RIO Platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT is currently supported by the CNRS, the Ministère de l'Éducation Nationale de l'Enseignement Supérieur et de la Recherche (MENESR), Université Montpellier 2 Plan Pluri-Formation, Région Languedoc-Roussillon, BIOTIC-LR2004, ACI-IMPBIO (IMP82-2004), GIS-AGENAE, Agence Nationale de la Recherche ANR (BIOSYS06_135457) and the European Union ImmunoGrid project (IST-2004-0280069).

References

- [1] M.-P. Lefranc, G. Lefranc, *The Immunoglobulin FactsBook*, Academic Press, London UK, 2001, 458 pp.
- [2] M.-P. Lefranc, G. Lefranc, *The T Cell Receptor FactsBook*, Academic Press, London UK, 2001, 398 pp.
- [3] H. Sakano, K. Huppi, G. Heinrich, S. Tonegawa, Sequences at the somatic recombination sites of immunoglobulin light-chain genes, *Nature* 280 (1979) 288–294.
- [4] M.-P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clément, D. Chaume, G. Lefranc, IMGT, the international ImMunoGeneTics information system[®], *Nucleic Acids Res.* 33 (2005) D593–D597.
- [5] M.-P. Lefranc, O. Clément, Q. Kaas, E. Duprat, P. Chastellan, I. Coelho, K. Combres, C. Ginestoux, V. Giudicelli, D. Chaume, G. Lefranc, IMGT-Choreography for immunogenetics and immunoinformatics, *In Silico Biol.* 5 (2005) 45–60.
- [6] V. Giudicelli, M.-P. Lefranc, *Ontology for Immunogenetics: IMGT-ONTOLOGY*, *Bioinformatics* 15 (1999) 1047–1054.
- [7] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, N. Bosc, G. Folch, D. Guiraudou, J. Jabado-Michaloud, S. Magris, D. Scaviner, V. Thouvenin, K. Combres, D. Girod, S. Jeanjean, C. Protat, M. Youssi Monod, E. Duprat, Q. Kaas, C. Pommié, D. Chaume, G. Lefranc, IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics <http://imgt.cines.fr>, *In Silico Biol.* 4 (2004) 17–29.
- [8] V. Giudicelli, D. Chaume, J. Jabado-Michaloud, M.-P. Lefranc, Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY, *Stud. Health Technol. Inform.* 116 (2005) 3–8.
- [9] Q. Kaas, M.-P. Lefranc, T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB, *In Silico Biol.* 5 (2005) 505–528.
- [10] V. Giudicelli, C. Ginestoux, G. Folch, J. Jabado-Michaloud, D. Chaume, M.-P. Lefranc, IMGT/LIGM-DB, the IMGT[®] comprehensive database of immunoglobulin and T cell receptor nucleotide sequences, *Nucleic Acids Res.* 34 (2006) D781–D784.
- [11] H.M. Wain, E.A. Bruford, R.C. Lovering, M.J. Lush, M.W. Wright, S. Povey, Guidelines for human gene nomenclature, *Genomics* 79 (2002) 464–470.
- [12] M.-P. Lefranc, Unique database numbering system for immunogenetic analysis, *Immunol. Today* 18 (1997) 509.
- [13] M.-P. Lefranc, The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains, *Immunologist* 7 (1999) 132–136.
- [14] M.-P. Lefranc, C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains, *Dev. Comp. Immunol.* 27 (2003) 55–77.
- [15] M.-P. Lefranc, C. Pommié, Q. Kaas, E. Duprat, N. Bosc, D. Guiraudou, C. Jean, M. Ruiz, I. Da Piedade, M. Rouard, E. Foulquier, V. Thouvenin, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains, *Dev. Comp. Immunol.* 29 (2005) 185–203.
- [16] M.-P. Lefranc, E. Duprat, Q. Kaas, M. Tranne, A. Thiriou, G. Lefranc, IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN, *Dev. Comp. Immunol.* 29 (2005) 917–938.
- [17] T.R. Gruber, A translation approach to portable ontologies, *Knowledge Acquisit.* 5 (1993) 199–220.
- [18] N. Guarino, P. Giaretta, Ontologies and knowledge bases: towards a terminological clarification, in: N. Mars (Ed.), *Towards Very Large Knowledge Bases*, IOS Press, Amsterdam, 1995, pp. 29–45.
- [19] N. Guarino, Understanding, building and using ontologies, *Int. J. Human-Comput. Stud.* 46 (1997) 293–310.
- [20] Noy N.F., McGuinness D.L., *Ontology development 101: A guide to creating your first ontology*, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

- [21] B. Smith, *Ontology*, in: L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information*, Blackwell, Oxford, 2003, pp. 155–166.
- [22] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, C. Rosse, *Relations in biomedical ontologies*, *Genome Biol.* 6 (2005) R46.
- [23] L.-N. Soldatova, A. Clare, A. Sparkes, R.D. King, *An ontology for a Robot Scientist*, *Bioinformatics* 22 (2006) e464–e471.
- [24] D. Chaume, V. Giudicelli, M.-P. Lefranc, *IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data*, in: *Proceedings of NETTAB 2001, Network Tools and Applications in Biology*, Genoa, Italy, May 17–18, 2001, pp. 71–75.
- [25] D. Chaume, K. Combres, V. Giudicelli, M.-P. Lefranc, *Retrieving factual data and documents using IMGT-ML in the IMGT information system[®]*, in: *Proceedings of NETTAB 2005, Workflows management: new abilities for the biological information overflow*, Naples, Italy, Oct 5–7, 2005, pp. 47–51.
- [26] N.F. Noy, M. Crubezy, R.W. Fergerson, H. Knublauch, S.W. Tu, J. Vendetti, et al., *Protege-2000: an open-source ontology-development and knowledge-acquisition environment*, *AMIA Annu. Symp. Proc.* (2003) 953.
- [27] J.T. Eppig, C.J. Bult, J.A. Kadin, J.E. Richardson, J.A. Blake, *the members of the Mouse Genome Database Group 2005, The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology*, *Nucleic Acids Res.* 33 (2005) D471–D475.
- [28] Q. Kaas, M. Ruiz, M.-P. Lefranc, *IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data*, *Nucleic Acids Res.* 32 (2004) D208–D210.
- [29] K. Eilbeck, S.E. Lewis, *Sequence Ontology annotation guide*, *Comp. Funct. Genomics* 5 (2004) 642–647.
- [30] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, *Entrez Gene: gene-centered information at NCBI*, *Nucleic Acids Res.* 35 (2007) D26–D31.
- [31] V. Giudicelli, D. Chaume, M.-P. Lefranc, *IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes*, *Nucleic Acids Res.* 33 (2005) D256–D261.
- [32] E. Duprat, Q. Kaas, V. Garelle, G. Lefranc, M.-P. Lefranc, *IMGT standardization for alleles and mutations of the V-LIKE-DOMAINS and C-LIKE-DOMAINS of the immunoglobulin superfamily*, in: S.G. Pandalai (Ed.), *Recent Research Developments in Human Genetics, 2*, Research Signpost, Trivandrum, India, 2004, pp. 111–136.
- [33] Q. Kaas, M.-P. Lefranc, *IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains*, *Curr. Bioinform.* 2 (2007) 21–30.
- [34] *The Gene Ontology Consortium, The Gene Ontology (GO) project in 2006*, *Nucleic Acids Res.* 34 (2006) D322–D326.
- [35] B. Peters, J. Sidney, P. Bourne, H.H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J.V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, A. Sette, *The immune epitope database and analysis resource: from vision to blueprint*, *PLoS Biol.* 3 (2005) e91.